

コード生成タスクにおける モデル崩壊の実証

S2310233 藤原颯太

背景：合成データの広がり

生成AIの普及

- ・ 大量の合成データがウェブ上に公開されるようになった

LLMはウェブ上のデータを収集して学習を行う

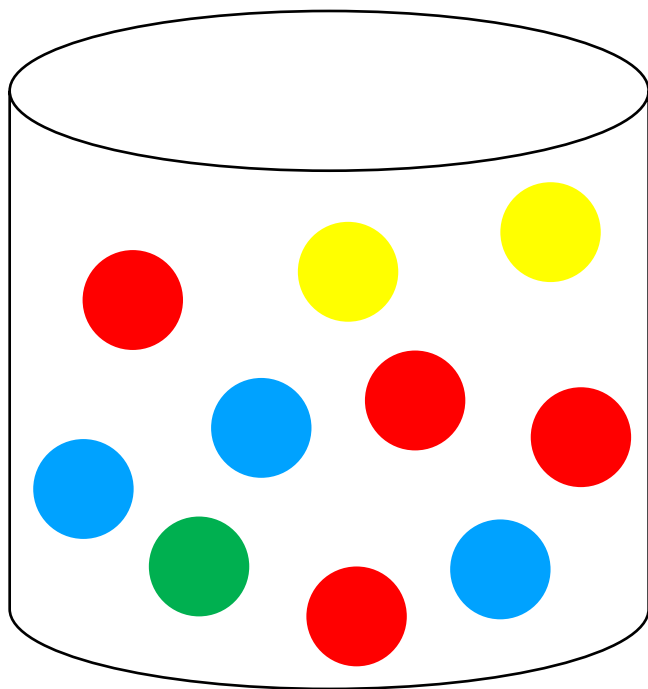


スクレイピングによる学習データの収集で
合成データを収集してしまう

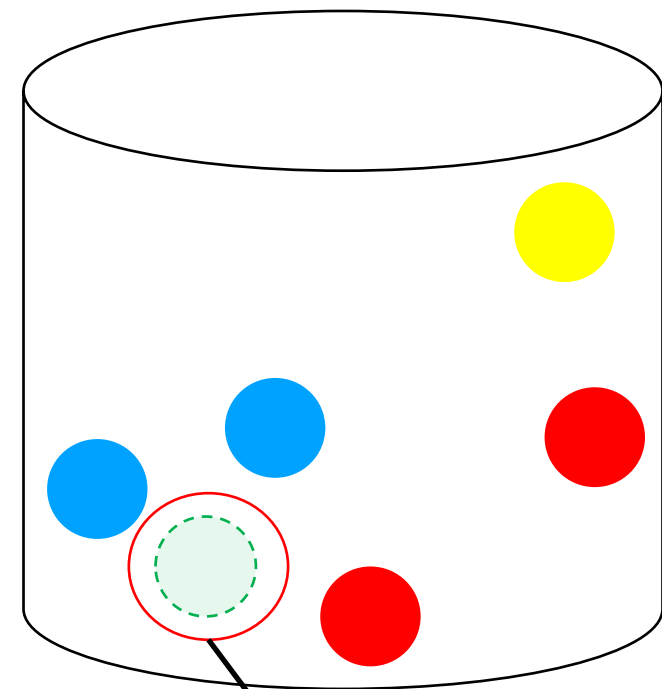
背景：合成データの問題点

有限サンプリングによる問題

- データを合成する際に元の分布に含まれる表現が失われる



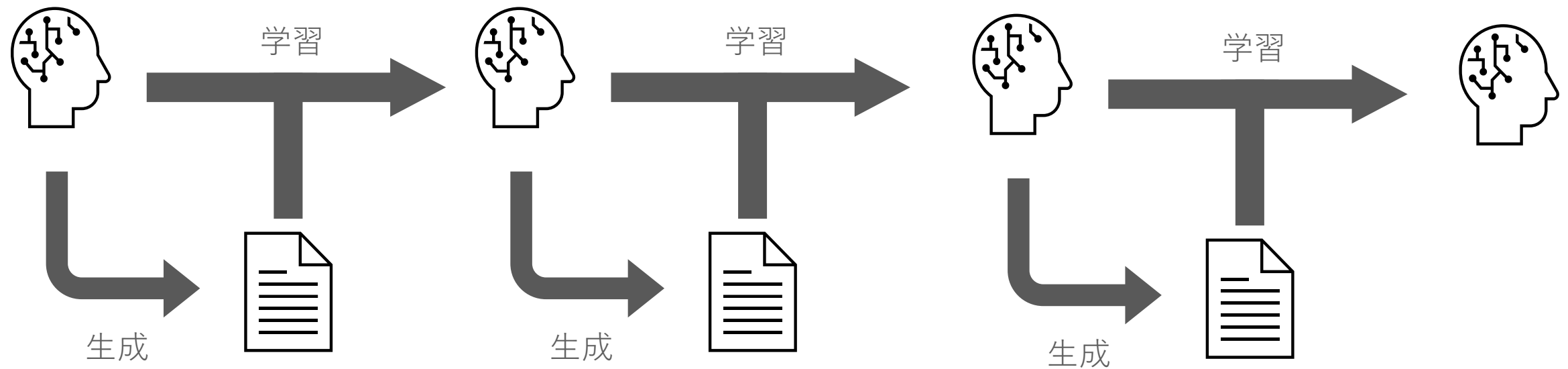
多様な表現の実データ



既存分布の表現を
損失した合成データ

背景：モデル崩壊とは

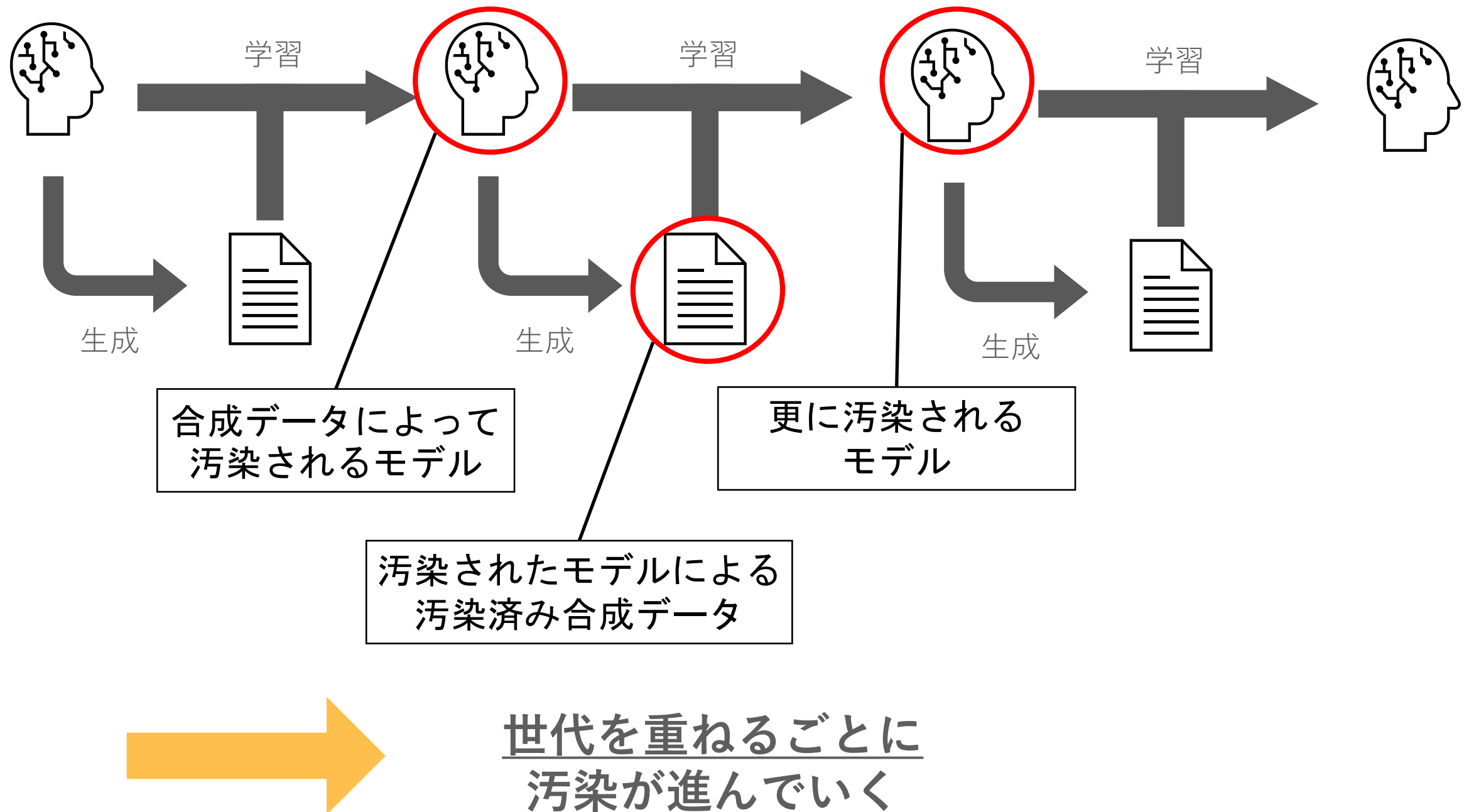
再帰的な反復学習による退行プロセス



- サンプルングによる誤差が蓄積
- 珍しい表現の情報が失われる
- 特定の表現が頻出するようになる

背景：モデル崩壊とは

再帰的な反復学習による退行プロセス



問題点

- | GitHub上のソースコードにも合成データが含まれるようになってきている
- | コード生成分野でのモデル崩壊現象の検証はされていない

目的・RQs

本研究の目的

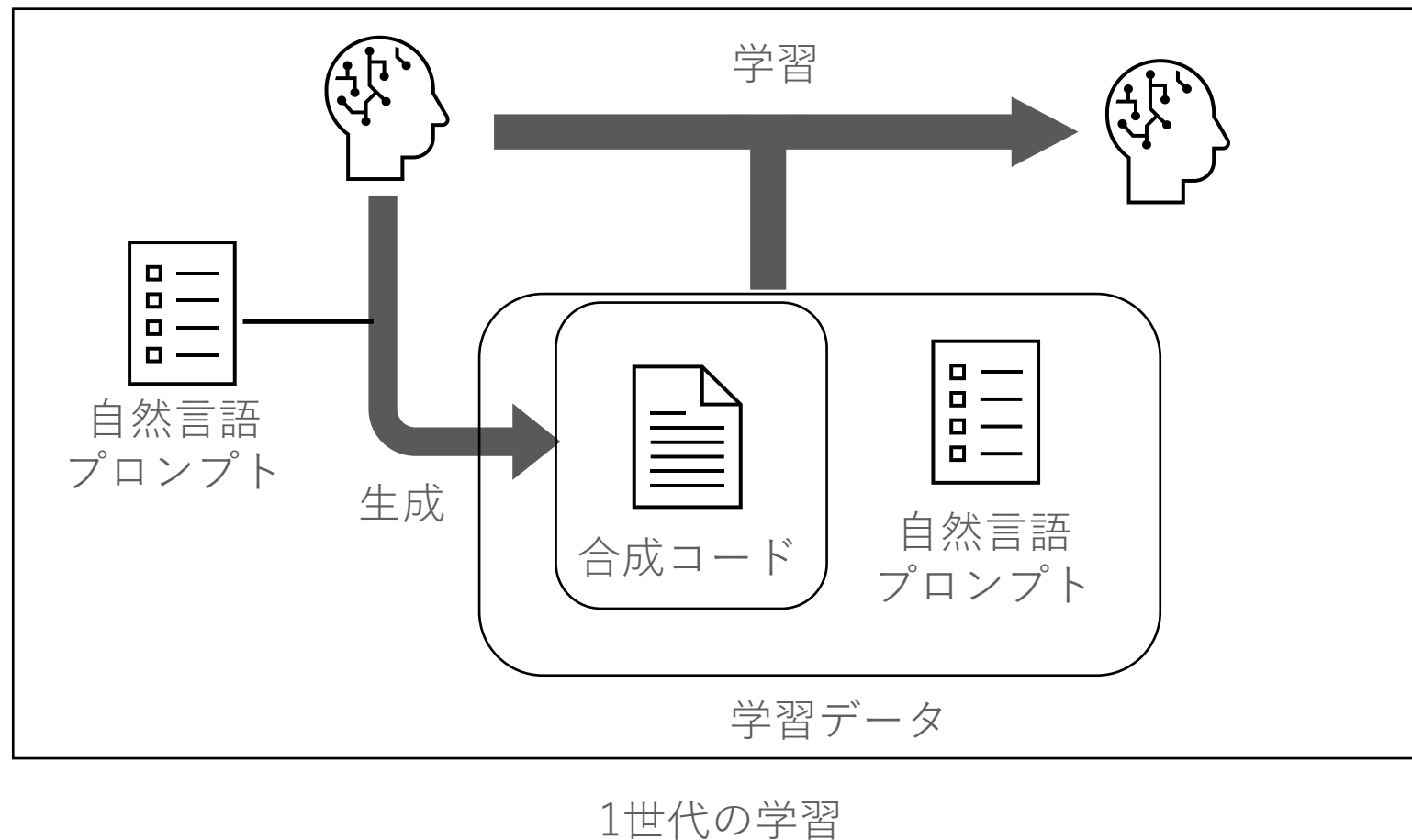
- コード生成タスクでのモデル崩壊現象を実証する

RQs

- RQ1
 - 再帰的な反復学習でコード生成モデルの性能は低下するか
- RQ2
 - 性能低下はモデル崩壊現象によるものか

実験概要

コード生成モデルで再帰的な反復学習を行う



- gen_0からgen_10まで学習を繰り返す

実験設定

モデル

- StarCoder2-3B
 - 学習に使われたデータが公開されている

データセット

- CodeSearchNet
 - 2019年に収集されたデータ
 - 実データのみで構成されている
 - StarCoder2-3B の学習データとの重複を除去して利用

実験設定

学習法

- QLoRA
 - 低ランク行列のアダプタを用いることで軽量化する手法
 - 学習ごとにアダプタを再学習する
- SFT
 - 教師あり学習によるファインチューニング
 - docstringを入力,コードを出力とする

合成データ生成

- temperature sampling
 - コード生成でよく用いられる手法
 - 入力プロンプトは固定

評価指標

ベンチマークによる評価(RQ1)

- モデルの性能を評価
 - HumanEval
 - MBPP+

PPLによる評価(RQ2)

- トークンの予測しやすさを評価する指標
 - gen_nモデルで実データのPPLを測定
 - gen_0モデルでgen_nモデルによる合成データのPPLを測定
- モデル崩壊による分布の変化を観測